REMARKS

Claims 1-20 are pending in the application. The pending claims are recited below for convenience.

1. (Previously amended) A computer-implemented method of determining content type of contents of a subject Web page, comprising the steps of:

providing a predefined set of potential content types;

for each potential content type, preparing a distinguishing series of tests, wherein the distinguishing series of tests includes at least one binary tests, at least one non-binary tests and at least one test: (a) examining syntax or grammar; or (b) examining page format or style other than position of data or a keyword in the subject Web page;

for each potential content type, running the distinguishing series of tests having test results which enable quantitative evaluation of at least some contents of the subject Web page being of the potential content type;

mathematically combining the test results; and

based on the combined test results, assigning a respective probability, for each potential content type, that some contents of that type exists on the subject Web page.

2. (Original) A method as claimed in Claim 1 wherein the set of potential content types include any combination of organization description, organization history, organization mission, organization products/services, organization members, organization contact information, management team information, job opportunities, press releases, calendar of events/activities, biographical data, articles/news with information about people, articles/news with information about organizations and employee roster.

3. (Original) A method as claimed in Claim 1 wherein the step of combining includes producing a respective confidence level for each potential content type, that at least some content of the subject Web page is of the potential content type.

4. (Original) A method as claimed in Claim 1 wherein the step of combining the test results includes using a Bayesian network.

5. (Original) A method as claimed in Claim 4 further comprising the step of training the Bayesian network using a training set of Web pages with respective known content types such that statistics on the test results are collected on the training set of Web pages.

6. (Previously amended) A method as claimed in Claim 1 wherein the predefined set includes a potential content type of press release and the distinguishing series of tests further includes at least one of:

(i) determining whether a predefined piece of data or keyword appears in the subject Web page;

(ii) examining text properties; and

(iii) determining whether the predefined piece of data or keyword appears in URLs in the subject Web page.

7. (Previously amended) The method as claimed in Claim 1 wherein the distinguishing series of tests further includes at least one of:

(i) determining whether a predefined piece of data or keyword appears in the subject Web page;

(ii) examining text properties; and

(iii) determining whether the predefined piece of data or keyword appears in URLs in the subject Web page.

8. (Original) A method as claimed in Claim 1 further comprising the step of storing indications of the assigned probabilities of each potential content type per respective Web page.

9. (Original) A database formed by the method of Claim 8, said database containing indications of Web pages and corresponding content types determined to be found on respective Web pages.

10. (Previously amended) Apparatus for determining content type of contents of a subject Web page, comprising:

a predefined set of potential content types, each potential content type being associated with a respective distinguishing series of tests, wherein the distinguishing series of tests includes at least one binary tests, at least one non-binary tests and at least one test: (a) examining syntax or grammar; or (b) examining page format or style order other than position of data or a keyword in the subject Web page;

a test module utilizing the predefined set, the test module employing the distinguishing series of tests as a plurality of processor-executed tests having test results which enable, for each potential content type, quantitative evaluation of at least some contents of the subject Web page being of the potential content type, for each potential content type, the test module (i) running the respective distinguishing series of tests, (ii) combining the test results and (iii) for each potential content type, assigning a respective probability that at least some contents of that type exists on the subject Web page being of the potential content type.

11. (Original) Apparatus as claimed in Claim 10 wherein the set of potential content types include any combination of contact information, press release, company description, employee list, other.

12. (Original) Apparatus as claimed in Claim 10 wherein the test module produces a respective confidence level for each potential content type, that at least some content of the subject Web page is of the potential content type.

13. (Original) Apparatus as claimed in Claim 10 wherein the test module combines the test results using a Bayesian network.

14. (Original) Apparatus as claimed in Claim 13 further comprising a training member for training the Bayesian network using a training set of Web pages with respective known content types, such that statistics on the test results are collected on the training set of Web pages.

15. (Original) Apparatus as claimed in Claim 10 wherein the predefined set includes a potential content type of at least one of organization description, organization history, organization mission, organization products/services, organization members, organization contact information, management team information, job opportunities, press releases, calendar of events/activities, biographical data, articles/news with information about people, articles/news with information about organizations and employee roster.

16. (Previously amended) Apparatus as claimed in Claim 15 wherein the processor-executed tests include at least one of:

(i) determining whether a predefined piece of data or keyword appears in the subject Web page;

(ii) examining text properties; and

(iii) determining whether the predefined piece of data or keyword appears in URLs in the subject Web page.

17. (Previously amended) Apparatus as claimed in Claim 10 wherein the processor-executed tests include any of:

(i) determining whether a predefined piece of data or keyword appears in the subject Web page;

(ii) examining text properties; and

(iii) determining whether the predefined piece of data or keyword appears in URLs in the subject Web page.

18. (Original) Apparatus as claimed in Claim 10 further comprising storage means for receiving and storing indications of the assigned probabilities of each content type per Web page as determined by the test module, such that the storage means provides a cross reference between a Web page and respective content types of contents found on that Web page.

19. (Previously presented) A method as claimed in Claim 1 wherein the at least one binary test and the at least one non-binary tests include one or more of the following tests:

i) whether the subject Web page contains a press release;

ii) whether the subject Web page has a title;

iii) whether the subject Web page has a copyright statement;

iv) whether the subject Web page has a navigation map;

v) whether the subject Web page has a line with a keyword followed by at least another keyword within the next 10, 20, 30 or 40 lines;

vii) whether a first sentence of a first paragraph of the subject Web page has a date;

viii) whether the first sentence of the first paragraph of the subject Web page is preceded by a header line;

ix) whether the first sentence of the first paragraph of the subject Web page contains the keyword or a form of the keyword;

xi) whether the subject Web page contains a text line starting with the keyword; and

xii) a calculation of a percentage of header lines, the average sentence length, number of different domains, number of lines that contain the keyword or number of phrases that contain the keyword.

20. (Previously presented) Apparatus as claimed in Claim 10 wherein the at least one binary test and the at least one non-binary tests include one or more of the following tests:

i) whether the subject Web page contains a press release;

ii) whether the subject Web page has a title;

iii) whether the subject Web page has a copyright statement;

iv) whether the subject Web page has a navigation map;

v) whether the subject Web page has a line with a keyword followed by at least another keyword within the next 10, 20, 30 or 40 lines;

vii) whether a first sentence of a first paragraph of the subject Web page has a date;

viii) whether the first sentence of the first paragraph of the subject Web page is preceded by a header line;

ix) whether the first sentence of the first paragraph of the subject Web page contains the keyword or a form of the keyword;

xi) whether the subject Web page contains a text line starting with the keyword; and

xii) calculation of a percentage of header lines, the average sentence length, number of different domains, number of lines that contain the keyword or number of phrases that contain the keyword.

35 U.S.C. § 103(a) Rejection of Claims 1-3, 6-9, 10-12 and 15-18 [*sic.*]

The present invention is directed to computer-implemented methods and apparatus for determining the content type of a subject Web page. Particularly, in base Claims 1 and 10, the invention includes a distinguishing series of tests. In addition to non-binary tests, the distinguishing series of tests include at least one test: (a) examining syntax or grammar; or (b) examining page format or style order other than the position of data or a keyword in the subject Web page. Examining syntax or grammar includes, for example, the number of passive sentences, number of sentences without a verb, percentage of verbs in past tenses, number of fonts used, existence of certain characters in determining the content type of the subject Web page. (*See* page 7, lines 8-11 and page 19, lines 4-12, Specification.)

Claims 1-3, 6-9, 10-12 and 15-18 have been rejected under 35 U.S.C. § 103(a) as being unpatentable over Russell-Falla et al. (U.S. Patent No. 6,675,162) (hereinafter "Russell-Falla") in view of Chakrabarti et al. (U.S. Patent No. 6,389,436) (hereinafter "Chakrabarti") in further view of van den Akker (U.S. Patent No. 6,415,250) (hereinafter "Akker").

The examiner stated that Akker teaches a test for classifying an incoming text, and the test includes probabilistic analysis of the inputted text which reflect morphological characteristics of natural languages, wherein the tests examine the syntax and grammar of the incoming text. According to the examiner, "it would have been obvious to one or ordinary skill in the art at the time of the invention for one of the distinguishing series of tests of Russell-Falla to have analyzed to syntax or grammar of the text because Akker teaches by analyzing the grammar or syntax of an incoming text to determine a source language provides the benefits of being able to efficiently and automatically classify and store documents as well as automatically selecting appropriate linguistic tools for the documents." (*See* in paragraph 7, pages 2-5, Office Action.) As such, Akker purportedly teaches the limitation of examining syntax or grammar in base claims 1 and 10, curing the deficiency by Russell-Falla and Chakrabarti to render the claims obvious.

Akker is directed to human language recognition technology. In particular, Akker is directed to an automatic language identification system for determining the source of language.

(*See* col. 3, lines 26-28 of Akker.) Embodiments of Akker identify the language (*i.e.* English, German, Dutch or French) in which a text is written based up on a morphological analysis of words that are contained in the text. (*See* col. 8, lines 58-60.) As shown in Figs. 2A, 2B and 2C of Akker, the analysis of identifying the language of a document requires examining several morphs or form of a word. For example, the analysis requires finding at least two different forms of the word "house" or "nice" to provide a probability score that the document containing either is in English. In another example, the disclosure of Akker analyzes suffixes of the predetermined groups of words to identify the language in which the text is written or inputted. (*See* col. 3, lines 13-23 and Figs. 2A, 2B and 2C of Akker.)

There is no motivation to combine the teaching of Akker with that of Chakrabarti to render the present invention recited by base claims 1 and 10 obvious

As stated earlier, it is the view of the PTO that Akker purportedly teaches the limitation of examining syntax or grammar and therefore, in view of Russell-Falla and in further view of Chakrabarti, the present invention recited in base claims 1 and 10 are obvious.

Applicants, however, respectfully disagree with the examiner because there is no motivation to combine the teaching of Akker with that of Chakrabarti. § 2141 of MPEP on "35 U.S.C. 103; the Graham factual inquiries" states the following:

> In determining the differences between the prior art and the claims, the question under 35 U.S.C. 103 is not whether the differences themselves would have been obvious, but whether the claimed invention as a whole would have been obvious. *Stratoflex, Inc. v. Aeroquip Corp.*, 713 F.2d 1530, 218 USPQ 871 (Fed. Cir. 1983); *Schenck v. Nortron Corp.*, 713 F.2d 782, 218 USPQ 698 (Fed. Cir. 1983)
>
> (*See* § 2141.02, I, 2100-122, left column, Original 8th Ed., August 2001, Latest Rev. Aug. 2006.)
>
> A prior art reference must be considered in its entirety, i.e., as a whole, including portions that would lead away from the claimed invention. *W.L. Gore & Associates, Inc. v. Garlock, Inc.*, 721 F.2d 1540, 220 USPQ 303 (Fed. Cir. 1983), *cert. denied*, 469 U.S. 851 (1984).
>
> (*See* § 2141.02, VI, 2100-124, left column, Original 8th Ed., August 2001, Latest Rev. Aug. 2006.)

Accordingly, when considered the entirety of both Chakrabarti and Akker, there is no suggestion or motivation in the references themselves to combine the reference teachings to render the present invention recited in base claims 1 and 10 obvious.

The invention taught in Chakrabarti is related to computer-implemented classifiers, and, in particular, to a hypertext classifier that classifies documents that contain hyperlinks. (*See* col. 1, lines 6-8 and col. 5, lines 46-53 of Chakrabarti.) Chakrabarti states the following:

> A text-based classifier classifies the documents based only on the text contained in the documents. However, *documents on the Web typically contain hyperlinks.* These hyperlinks are ignored by text-based classifiers, although the hyperlinks contain useful information for classification...*There is a need in the art for an improved classifier that can classify documents containing hyperlinks*.... Unlike conventional systems, the hypertext classifier 110 of the present invention exploits *topic information* implicitly present in hyperlink structure..
>
> (*See* col. 2, lines 29-34 col. 3, lines 63-64 and col. 7 lines 5-7 of Chakrabarti; italics added)

As the title of Chakrabarti, which is ENHANCED HYPERTEXT CATEGORIZATION USING HYPERLINKS, indicates, the system of Chakrabarti uses the hyperlinks to classify a document into categories such as "news, entertainment, sports, business or theater." (*See* col. 6, lines 62-64 of Chakrabarti.) Therefore, for the system of Chakrabarti, the analysis starts with examining hypertexts such as "Reuters at http://www.research.att.com/lewis" or "MEDLINE at http://medir.ohsu.edu/pub/ohsumed".

However, while Akker supposedly cure the deficiencies suffered by Russell-Falla and Chakrabarti in lacking the limitation of examining syntax or grammar, Chakrabarti teaches away from incorporating the teachings of Akker. Contrary to the contents in a typical document analyzed by the language identifier taught Akker, hyperlinks in a web page are not expressed in morphs or different forms. One skilled in the art would not be motivated to combine the teachings of Chakrabarti and Akker to examine syntax or grammar of a document. Considering references as a whole as directed by MPEP, such combination would change the principle of operation of the invention taught in Chakrabarti. The characters that make up a hyperlink typically analyzed by the invention in Chakrabarti is so restrictive that even a minute typographical error (*i.e.* switching a "/" (forward slash) to a "-" (a dash)) of a hyperlink can lead

to an error web page or a wrong web page. As such, a hyperlink must be expressed precisely and a web pages to be classified in Chakrabarti must contain only one form rather than further having other morphs of the examples above such as "Reuter at http://www.research.att.com/lewises" or "MEDLINAL at http://medir.ohsu.edu/public/ohsumed" (changes from the earlier examples are highlighted). Therefore, one skilled in the art would not apply the language identifier of Akker that specifically search for morphs of a word to classify the content of a document in Chakrabarti.

Furthermore, because the text-base classification of Chakrabarti is limited to a rudimentary analysis, it discounts the possibility of a sophisticated linguistic analysis such as one of Akker:

> In general, *beyond elementary case-conversion*, stop word filtering, and stemming, *the text-based classifier performs no linguistic operations*, letting only the feature statistics drive the technique. This makes the system more efficient and robust.
>
> (*See* col. 11. lines 3-7 of Chakrabarti; italics and underlining added.)

As such, Chakrabarti further teaches away from combining the teaching of Akker to render the present invention recited in base claims 1 and 10 obvious.

In view of foregoing, Applicants respectfully request the § 103(a) rejection of base claims 1 and 10 be withdrawn. Furthermore, as claims 2, 6-9, 11-12 and 15-18 depend from Claim 1 or 10, these claims are allowable for the same reasons discussed above.

Although the examiner did not specifically include claims 19 and 20 on paragraph 7, page 2 of the Office Action as rejected claims, he discussed the reasons for rejecting these claims on page 6. Therefore, it appears that the examiner mistakenly excluded claims 19 and 20 initially from the rest of the claims 1-3, 6-9, 10-12 and 15-18 that stand rejected under 35 U.S.C. § 103(a). Nonetheless, as claims 19 and 20 are also dependent from claims 1 and 10, respectively, these claims are also allowable for the same reasons as claims 1 and 10.

35 U.S.C. § 103(a) Rejection of Claims 4, 5, 13 and 14

Claims 4, 5, 13 and 14 have been rejected under 35 U.S.C. § 103(a) as being unpatentable over Russell-Falla in view of Chakrabarti in view of Akker and in further view of Haug et al. (U.S. Patent No. 6,556,964) (hereinafter "Haug"). Haug is directed to a probabilistic model for determining the meaning of sentences or phrases in medical reports. The Haug model extracts and encodes medical concepts using a Bayesian network.

Claims 4-5, 13 and 14 depend from base Claims 1 or 10. Therefore, Claims 4, 5, 13 and 14 also include the element of the distinguishing series of tests having both binary and non-binary tests. As explained above, neither Russell-Falla, Chakrabarti nor Akker teaches, suggests or otherwise makes obvious the distinguishing series of tests having one or more tests involving syntax, grammar or page style of Claims 4, 5, 13 and 14. Furthermore, Haug does not cure this deficiency to make Claims 4, 5, 13 and 14 obvious. Therefore, Applicants respectfully request the § 103(a) rejection of Claims 4, 5, 13 and 14 be withdrawn.

CONCLUSION

In view of the above remarks, it is believed that all claims (claims 1-20) are in condition for allowance, and it is respectfully requested that the application be passed to issue. If the Examiner feels that a telephone conference would expedite prosecution of this case, the Examiner is invited to call the undersigned.

Respectfully submitted,

HAMILTON, BROOK, SMITH & REYNOLDS, P.C.

By _____
H. Joon Chung
Registration No. 52,748
Telephone: (978) 341-0036
Facsimile: (978) 341-0136

Concord, MA 01742-9133
Dated: 12/18/06